•Biostatistics in psychiatry (17)•

From pilot studies to confirmatory studies

Naihua DUAN

1. Introduction

Pilot studies serve an important role in clinical research. They provide information needed to design large confirmatory studies that aim to provide definitive evidence on the efficacy/effectiveness of novel interventions. In the drug development process, phase I and II studies are used to prepare for the conduct of confirmatory phase III trials. In investigator-initiated studies supported by the government such a long developmental process is usually unfeasible, so pilot studies for these projects typically use smaller samples over shorter periods of time than planned for the main study. As a planning tool, pilot studies can be used for a variety of objectives. First, pilot studies can be used to assess the feasibility of confirmatory studies, to rule out those that might sound appealing on paper but cannot be implemented effectively in practice. Second, pilot studies can be used to 'debug' the protocols of confirmatory studies, to help identify remedies necessary to ensure the success for the subsequent confirmatory studies. Third, pilot studies can be used to guide the prioritization of resource allocation, to pursue the most promising interventions, and to withhold further efforts from not-so-promising interventions.

It should be noted that a pilot study cannot be a stand-alone study. Rather, it is necessary for a pilot study to be designed and evaluated along with its companion confirmatory study, which is to be conducted after the pilot study if the findings from the pilot study support the decision to move forward with the confirmatory study. As a planning tool for the confirmatory study, the pilot study should be designed specifically to address the needs of the confirmatory study. Therefore, I assume for the rest of this essay that the pilot study and the confirmatory study are being considered jointly.

The framing of pilot studies has received extensive discussion in recent mental health statistics literature,[1,2] leading to major additions to the directives provided in several recent program announcements from the U.S. National Institute of Mental Health (NIMH).[3-8] These new directives indicate that the aims of pilot studies should be on the feasibility of the proposed confirmatory study, rather than on the preliminary estimation of the effect size for the novel intervention that will be assessed in the confirmatory study.

Up until a few years ago, the prevailing way to use pilot studies was as miniaturized versions of the confirmatory studies under consideration. Pilot studies used the same randomized design as the confirmatory study but with much smaller sample sizes; the goal was to produce preliminary estimates for the effect sizes of the novel interventions under investigation and then to use these estimates to prioritize resource allocation. Consider, for example, ten candidate interventions that appear to be promising. Under this 'test and select' paradigm, the research program sponsors ten pilot studies, then selects the three interventions with the largest preliminary effect sizes to conduct confirmative studies, dropping the remaining seven candidate interventions from further consideration.

Although this test and select procedure appears to be reasonable and has been used widely for a number of years, it may not result in the selection of the most promising interventions. There is a phenomenon that has become known as 'regression towards the mean[9] for R01s' in which the finalist interventions that appeared to be most promising during the pilot phase often perform less well during the confirmatory phase ('R01' is the label for confirmatory studies funded by the U.S. National Institute of Health).[10] This counter-intuitive result is a product of the selection procedure used to identify the most promising interventions. In the preceding example the reason the three finalists were selected out of the pool of ten candidate interventions is usually a combination of true superiority and chance. If the pilot studies were conducted with large sample sizes to reduce the influence of chance, this selection procedure would be more likely to reflect true superiority of the interventions. However, pilot studies are usually conducted with small sample sizes; in such studies the signal-to-noise ratio is usually low so the relative importance of chance in the final outcome (i.e., effect size) versus that of the true superiority of the intervention is large. Thus the three interventions with the largest effect sizes in the pilot studies might not be the best interventions; during the confirmatory phase it

doi: 10.3969/j.issn.1002-0829.2013.05.011

Department of Psychiatry, Columbia University, New York City, USA
*correspondence: naihua.duan@columbia.edu

is likely that their effect sizes would regress toward the mean effect size of all ten interventions.

These problems with the test and select procedure were raised by Kraemer and colleagues [1] and discussed further by Leon and colleagues,[2] leading to the promulgation of a series of directives in several recent program announcements from the U.S. NIMH.[3-8] For example, NIMH's recent program announcement, PAR-12-279, entitled 'Pilot Intervention and Services Research Grants (R34)', invites investigators to pursue

> "research on 1) the development and/or pilot testing of novel or adapted interventions, 2) the adaptation and/or pilot testing of interventions with demonstrated efficacy for use in broader scale effectiveness trials, or 3) innovative services research directions that require preliminary testing or development. The R34 award mechanism provides resources for evaluating the feasibility, tolerability, acceptability, and safety of novel approaches to improving mental health and modifying health risk behavior, and for obtaining the preliminary data needed as a prerequisite to a larger-scale intervention (efficacy or effectiveness) or services study."[4]

Within this framework,

> "**Pilot intervention studies… do not necessarily need to be scaled down randomized controlled trials (RCTs) that propose formal tests of intervention outcomes.** Indeed, as described below, depending on the stage of intervention development, the R34 project might not involve randomization, but rather might focus on earlier stages such as the operationalization of an intervention protocol and corresponding manual and/or pilot testing of the experimental intervention in a case series with a sample drawn from the target population. **Most importantly, the R34 should propose the developmental work to be performed that would enhance the probability of success in a larger trial.** This is best done by working out the details of the experimental protocols, including the assessment protocol, the experimental intervention protocol, as well as the comparison intervention protocol and randomization procedures (if appropriate); examining the feasibility of recruiting and retaining participants into the study conditions (including the experimental condition and the comparison condition, if relevant); and developing supportive materials and resources." [4] (Text was transcribed from the original document, including the use of boldface font for emphasis.)

The program announcement further declares:

> "…collection of preliminary data regarding feasibility, acceptability, safety, tolerability, and target outcomes is appropriate. **However, given the intended pilot nature of the R34 mechanism, conducting formal tests of outcomes or attempting to obtain an estimate of an effect size is often not justified.** Given the limited sample sizes typically supportable under this pilot study mechanism, the variability in the effect sizes obtained is often so large as to be unreliable. Thus, using these potentially unstable effect size estimates in power calculations for larger studies, without regard to clinical meaningfulness, is not advisable."[4]

Similar statements were made in a number of other NIMH program announcements, including PAR-09-173[3] (an earlier version of PAR-12-279), PAR-13-188,[5] RFA-MH-14-101,[6] RFA-MH-14-102,[7] and RFA-MH-14-212.[8]

Two principles underlie these NIMH program announcements: pilot studies should focus on ascertaining the feasibility, tolerability, acceptability, and safety of novel interventions; and de-emphasis on the previous use of pilot studies to provide preliminary estimates for effect sizes. This change in the purpose of pilot studies raises several important statistical issues, two of which will be discussed below.

2. Use of comparison group(s) and randomization

The NIMH program announcements leave the options open for investigators to determine whether their pilot studies should be designed with a comparison group. For some pilot studies that are focused entirely on the feasibility of the novel intervention, it might be reasonable to deliver the novel intervention to all study participants, without a comparison group, to maximize the study's ability to ascertain feasibility and debug potential problems for the novel intervention. However, some pilot studies might also need to ascertain the feasibility of the randomized design to ensure that future candidate participants for the confirmatory study will be amenable to be randomized into either the experimental group or the comparison group. Indeed, the confirmatory study might fail if the vast majority of eligible participants have a strong preference for the experimental condition, violating the ethical requirement of equipoise for randomized assignment (i.e., clinical investigators should not force participants into a treatment condition that they prefer not to be assigned to). Under those circumstances, it would be necessary for the pilot study to be designed with both an experimental group and a comparison group, with a protocol to randomize participants between the two groups in order to ascertain the feasibility for the randomized design for the future confirmatory study.

As a side note, assessment of the feasibility of a randomized design might not require actual pilot testing with eligible participants. Rather, it might be possible to use appropriate survey techniques to inquire about the behavioral intention among eligible participants. (With questions phrased as follows: "If such a study were presented to you, would you be willing to be randomized to either the experimental group or the comparison group, or would you have a strong preference to be assigned to the experimental group and reject the comparison condition?") Although the behavioral intention elicited through such a survey might not predict actual participant behavior perfectly, the correlation between stated behavioral intention and actual behavior might be sufficiently strong to identify a potential problem. If a large majority of eligible participants reject the randomized design in such a survey, the confirmatory study is likely to run into problems with recruitment.

3. Skewed randomization

If a researcher decides to randomize subjects in the pilot study in order to ascertain the feasibility of using a randomized design in the confirmatory study, using the usual 1:1 randomization scheme (i.e., with equal probability for each participant to be randomized to the experimental or comparison groups) might not be appropriate. Such pilot studies aim both to assess the feasibility of a randomized design and to assess the feasibility of the novel intervention. The comparison group informs the first objective on the feasibility of the randomized design, but does not inform the second objective on the feasibility of the novel intervention. The experimental group informs both objectives. Therefore it is reasonable to allocate a larger share of the pilot study sample into the experimental group, to adequately fulfill the second objective of assessing the feasibility of the novel intervention. The extent to which the randomization should be skewed towards the experimental group depends on the relative importance between the two objectives. If the first objective on the feasibility of the randomized design is the over-riding objective, the randomization should not be skewed at all, to provide the best information possible on this objective. On the other hand, if the second objective on the feasibility of the novel intervention is the over-riding objective, the randomization should be skewed heavily towards the experimental group, to maximize the information available on the second objective.

In order to balance the needs for the two competing study objectives, a formal statistical decision for the randomization ratio can be made applying the weighted A-optimality criterion,[11,12] which uses weights that reflect the relative importance between the two study objectives.

4. Sample size determination

For the objectives of ascertaining feasibility and debugging potential problems, it is important to conduct power analyses to ensure that the pilot study has an adequate sample size to inform the design decisions that need to be made prior to conducting the confirmatory study.

We describe below the power analyses for these objectives, based on the binomial distribution for the probability of occurrence of specific problems with feasibility, denoted by *p*; examples would include the probability that eligible study participants would refuse participation in the comparison group, or the probability that a participant drops out from the novel intervention due to side-effects. The null hypothesis is that there are no feasibility problems, that is, p=0; the alternative is that p>0. As an example, consider a pilot study designed to assign n participants to the experimental group. Is this sample adequate to detect major problems with the novel intervention protocol? In order to conduct this power analysis, the investigators need to specify a threshold for the prevalence of the potential problem to be detected under the alternative hypothesis. Let us assume that the threshold is p=30%; in this case the pilot study would aim to detect major protocol problems that affect at least 30% of study participants, so any protocol problem with a prevalence of less than 30% is considered unimportant and would not necessarily be detected.

Let *k* denote the number of study participants who manifest the protocol problem of interest. We define the decision rule as follows:

Protocol problem is detected if k>0; and

No protocol problem is detected if k=0.

We define the power for the procedure to be the probability to detect a protocol problem:

$$\text{Power}=P\ (k>0).$$

Assuming that the protocol problem affects *n* study participants independently, the power defined above is given as follows:

$$\text{Power}=P\ (k>0)=1–(1–p)^n,$$

where *n* denotes the total number of participants exposed to the protocol, *k* denotes the number of participants who manifest the protocol problem, and *p* denotes the detection threshold for the protocol problem.

If we assume the detection threshold to be p=30%, the power defined above is then given as follows:

$$\text{Power}=P\ (k>0)=1–(1–p)^n=1– (1–0.3)^n=1–0.7^n.$$

Assume further that the study desires to have power of at least 80% to detect protocol problems with prevalence of 30% or higher. The power defined above is then given as follows:

$$0.8=1–0.7^n.$$

The sample size required, *n*, can be solved as follows:

$$0.7^n=1–0.8=0.2;$$

$$n \log(0.7)=\log(0.2);$$

$$n=\log(0.2)/\log(0.7)=4.51$$

Therefore, a sample size of five participants is adequate to provide more than 80% power to detect protocol problems with prevalence of 30% or more. (Plugging this sample size into the earlier expression for power, the power actual achieved is 83%, more than the target power of 80%.)

5. Conclusions

The conceptualization of pilot studies as planning tools focused on feasibility and debugging, rather than on the estimation of effect sizes, provides opportunities for innovative approaches to designing pilot studies that provide useful information for the subsequent confirmatory studies. Further collaborations between biostatisticians and clinical investigators in this area are warranted to develop an effective and seamless interface between pilot studies and confirmatory studies.

Conflict of interest

The author reports no conflict of interest related to this manuscript.

References

1. Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA. Caution regarding the use of pilot studies to guide power calculations for study proposals. *Arch Gen Psychiatry* 2006; **63**(5): 484-489.
2. Leon AC, Davis LL, Kraemer HC. The role and interpretation of pilot studies in clinical research. *J Psychiatr Res* 2011; **45**(5): 626-629. Epub 2010 Oct 28.
3. U.S. National Institute of Mental Health (NIMH) [Internet]. Bethesda (MD): National Institutes of Health (NIH) [updated 2009 Apr 22; cited 2013 Aug 19]. PAR-09-173: Pilot Intervention and Services Research Grants (R34). Available from: http://grants.nih.gov/grants/guide/pa-files/PAR-09-173.html.
4. U.S. National Institute of Mental Health (NIMH) [Internet]. Bethesda (MD): National Institutes of Health (NIH) [updated 2012 Sep 7; cited 2013 Aug 19]. PAR-12-279: Pilot Intervention and Services Research Grants (R34). Available from: http://grants.nih.gov/grants/guide/pa-files/PAR-12-279.html.
5. U.S. National Institute of Mental Health (NIMH) [Internet]. Bethesda (MD): National Institutes of Health (NIH) [updated 2013 Apr 10; cited 2013 Aug 19].PAR-13-188: Reducing the Duration of Untreated Psychosis in the United States (R34). Available from: http://grants.nih.gov/grants/guide/pa-files/PAR-13-188.html.
6. U.S. National Institute of Mental Health (NIMH) [Internet]. Bethesda (MD): National Institutes of Health (NIH) [updated 2013 May 30; cited 2013 Aug 19]. RFA-MH-14-101: Services Research for Autism Spectrum Disorder across the Lifespan (ServASD): Pilot Research on Services for Transition-Age Youth (R34). Available from: http://grants.nih.gov/grants/guide/rfa-files/RFA-MH-14-101.html.
7. U.S. National Institute of Mental Health (NIMH) [Internet]. Bethesda (MD): National Institutes of Health (NIH) [updated 2013 May 30; cited 2013 Aug 19]. RFA-MH-14-102: Services Research for Autism Spectrum Disorders across the Lifespan (ServASD): Pilot Studies of Services Strategies for Adults with ASD (R34). Available from: http://grants.nih.gov/grants/guide/rfa-files/RFA-MH-14-102.html.
8. U.S. National Institute of Mental Health (NIMH) [Internet]. Bethesda (MD): National Institutes of Health (NIH) [updated 2013 Aug 7; cited 2013 Aug 19]. RFA-MH-14-212: Research to Improve the Care of Persons at Clinical High Risk for Psychotic Disorders (R34). Available from: http://grants.nih.gov/grants/guide/rfa-files/RFA-MH-14-212.html.
9. Bland JM, Altman DG. Regression towards the mean. *BMJ* 1994; **308**(6942): 1499.
10. U.S. National Institute of Health (NIH) [Internet]. Bethesda (MD): National Institutes of Health (NIH) [updated 2013 Sep 12; cited 2013 Sept 28]. Grants and Funding/Types of Grant Programs/Research Grants. Available from: http://grants.nih.gov/grants/funding/funding_program.htm#RSeries.
11. Frankel MR, Shapiro MF, Duan N, Morton SC, Berry SH, Brown JA, et al. National probability samples in studies of low-prevalence diseases. Part II: Designing and implementing the HIV Cost and Services Utilization Study sample. *Health Serv Res* 1999; **34**(5): 969-992.
12. Shirakura T, Tong W-P. Weighted A-optimality for fractional 2^m factorial designs of resolution *V*. *J Stat Plan Inference* 1996; **56**(2): 243-256.

Professor Naihua Duan retired in 2012 from the Department of Psychiatry at Columbia University in New York where he served as a tenured Professor of Biostatistics (in Psychiatry) and as Director of the Division of Biostatistics. Professor Duan served as the Statistical Editor for the Shanghai Archives of Psychiatry from 2011 to 2012. His research interests include health services research, prevention research, sample design and experimental design, model robustness, transformation models, multilevel modeling, nonparametric and semi-parametric regression methods, and environmental exposure assessment.